Black Box Variational Inference with a Deterministic Objective

Faster, More Accurate, and Even More Black Box

Giordano, Ryan [1] Ingram, Martin[3] Broderick, Tamara[2]

September 7th, 2023

[1]University of California, Berkeley

[2]Massachusetts Institute of Technology

[3]University of Melbourne, Australia

Problem statement

We all want to do accurate Bayesian inference quickly:

- In terms of compute (wall time, model evaluations, parallelism)
- In terms of analyst effort (tuning, algorithmic complexity)

Markov Chain Monte Carlo (MCMC) can be straightforward and accurate but slow.

Problem statement

We all want to do accurate Bayesian inference quickly:

- In terms of compute (wall time, model evaluations, parallelism)
- In terms of analyst effort (tuning, algorithmic complexity)

Markov Chain Monte Carlo (MCMC) can be straightforward and accurate but slow.

Black Box Variational Inference (BBVI) can be faster alternative to MCMC. **But...**

- BBVI is cast as an optimization problem with an intractable objective $\Rightarrow$
- Most BBVI methods use **stochastic gradient** (SG) optimization $\Rightarrow$
  - SG algorithms can be hard to tune
  - Assessing convergence and stochastic error can be difficult
  - SG optimization can perform worse than second-order methods on tractable objectives
- Many BBVI methods employ a **mean-field (MF) approximation** $\Rightarrow$
  - Posterior variances are poorly estimated

Problem statement

We all want to do accurate Bayesian inference quickly:

- In terms of compute (wall time, model evaluations, parallelism)
- In terms of analyst effort (tuning, algorithmic complexity)

Markov Chain Monte Carlo (MCMC) can be straightforward and accurate but slow.

Black Box Variational Inference (BBVI) can be faster alternative to MCMC. **But...**

- BBVI is cast as an optimization problem with an intractable objective $\Rightarrow$
- Most BBVI methods use **stochastic gradient** (SG) optimization $\Rightarrow$
  - SG algorithms can be hard to tune
  - Assessing convergence and stochastic error can be difficult
  - SG optimization can perform worse than second-order methods on tractable objectives
- Many BBVI methods employ a **mean-field (MF) approximation** $\Rightarrow$
  - Posterior variances are poorly estimated

Our proposal: replace the intractable BBVI objective with a fixed approximation.

- Better optimization methods can be used (e.g. true second-order methods)
- Convergence and approximation error can be assessed directly
- Can correct posterior covariances with linear response covariances
- This technique is well-studied (but there's still work to do in the context of BBVI)

$\Rightarrow$ **Simpler, faster, and better BBVI posterior approximations ... in some cases.**

Outline

- BBVI Background and our proposal
  - Automatic differentiation variational inference (ADVI) (a BBVI method)
  - Our approximation: "Deterministic ADVI" (DADVI)
  - Linear response (LR) covariances
  - Estimating approximation error
- Experimental results: DADVI vs ADVI
  - DADVI converges faster than ADVI, and requires no tuning
  - DADVI's posterior mean estimates' accuracy are comparable to ADVI
  - DADVI+LR provides more accurate posterior variance estimates than ADVI
  - DADVI provides accurate estimates of its own approximation error
  - But stochastic ADVI often results in better objective function values (eventually)
- Theory and shortcomings
  - Pessimistic dimension dependence results from optimization theory
  - ...which do not apply in certain BBVI settings.
  - DADVI fails for expressive BBVI approximations (e.g. full-rank ADVI)
  - More work to be done!

Notation

$$\begin{aligned} \text{Data:} \quad & y \\ \text{Likelihood:} \quad & \mathcal{P}(y|\theta) \\ \text{Parameter:} \quad & \theta \in \mathbb{R}^{D_\theta} \\ \text{Prior:} \quad & \mathcal{P}(\theta) \quad \text{(density w.r.t. Lebesgue } \mathbb{R}^{D_\theta}\text{, nonzero everywhere)} \end{aligned}$$

We will be interested in means and covariances of the posterior $\mathcal{P}(\theta|y)$.

Notation

$$\begin{aligned} \text{Data:} \quad & y \\ \text{Likelihood:} \quad & \mathcal{P}(y|\theta) \\ \text{Parameter:} \quad & \theta \in \mathbb{R}^{D_\theta} \\ \text{Prior:} \quad & \mathcal{P}(\theta) \quad \text{(density w.r.t. Lebesgue } \mathbb{R}^{D_\theta}\text{, nonzero everywhere)} \end{aligned}$$

We will be interested in means and covariances of the posterior $\mathcal{P}(\theta|y)$.

,

Example: Election modeling (2016 US POTUS)

Data y :	Polling data (colored dots)
Likelihood $\mathcal{P}(y\|\theta)$:	Time series with random effects
Parameter θ :	15, 098-dimensional
Interested in:	Vote share on election day
MCMC time:	643 minutes (PyMC3 NUTS)

How can we approximate the posterior more quickly?
One answer: variational inference.

Variational inference [Blei et al., 2016]

We want the posterior $\mathcal{P}(\theta|y)$. Let $\mathrm{KL}\left(\mathcal{Q}(\theta)||\mathcal{P}(\theta)\right)$ denote KL divergence:

$$\mathrm{KL}\left(\mathcal{Q}(\theta)||\mathcal{P}(\theta)\right) = \mathop{\mathbb{E}}_{\mathcal{Q}(\theta)}\left[\log \mathcal{Q}(\theta)\right] - \mathop{\mathbb{E}}_{\mathcal{Q}(\theta)}\left[\log \mathcal{P}(\theta)\right].$$

The KL divergence is zero if and only if the two distributions are the same.

Variational inference [Blei et al., 2016]

We want the posterior $\mathcal{P}(\theta|y)$. Let $\mathrm{KL}\left(\mathcal{Q}(\theta)||\mathcal{P}(\theta)\right)$ denote KL divergence:

$$\mathrm{KL}\left(\mathcal{Q}(\theta)||\mathcal{P}(\theta)\right) = \underset{\mathcal{Q}(\theta)}{\mathbb{E}}\left[\log \mathcal{Q}(\theta)\right] - \underset{\mathcal{Q}(\theta)}{\mathbb{E}}\left[\log \mathcal{P}(\theta)\right].$$

The KL divergence is zero if and only if the two distributions are the same.

$$\text{A tautology:} \qquad \mathcal{P}(\theta|y) = \underset{\mathcal{Q}}{\operatorname{argmin}} \, \mathrm{KL}\left(\mathcal{Q}(\theta)||\mathcal{P}(\theta|y)\right)$$

$$\text{Variational inference:} \qquad \overset{*}{\mathcal{Q}}(\theta) = \underset{\mathcal{Q}\in\Omega_{\mathcal{Q}}}{\operatorname{argmin}} \, \mathrm{KL}\left(\mathcal{Q}(\theta)||\mathcal{P}(\theta|y)\right) \qquad \text{... for restricted } \Omega_{\mathcal{Q}}$$

Variational inference [Blei et al., 2016]

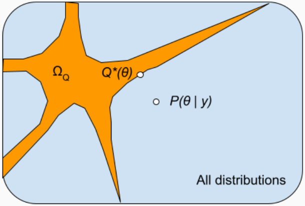
We want the posterior $\mathcal{P}(\theta|y)$. Let $\mathrm{KL}\left(\mathcal{Q}(\theta)||\mathcal{P}(\theta)\right)$ denote KL divergence:

$$\mathrm{KL}\left(\mathcal{Q}(\theta)||\mathcal{P}(\theta)\right) = \underset{\mathcal{Q}(\theta)}{\mathbb{E}}\left[\log \mathcal{Q}(\theta)\right] - \underset{\mathcal{Q}(\theta)}{\mathbb{E}}\left[\log \mathcal{P}(\theta)\right].$$

The KL divergence is zero if and only if the two distributions are the same.

$$\text{A tautology:} \qquad \mathcal{P}(\theta|y) = \underset{\mathcal{Q}}{\operatorname{argmin}}\, \mathrm{KL}\left(\mathcal{Q}(\theta)||\mathcal{P}(\theta|y)\right)$$

$$\text{Variational inference:} \qquad \overset{*}{\mathcal{Q}}(\theta) = \underset{\mathcal{Q}\in\Omega_{\mathcal{Q}}}{\operatorname{argmin}}\, \mathrm{KL}\left(\mathcal{Q}(\theta)||\mathcal{P}(\theta|y)\right) \qquad \text{... for restricted } \Omega_{\mathcal{Q}}$$



We hope to choose $\Omega_{\mathcal{Q}}$ so that

- The optimization problem is tractable
  - $\rightarrow$ simple $\Omega_{\mathcal{Q}}$ are better
- The best approximation is a good one
  - $\rightarrow$ complex $\Omega_{\mathcal{Q}}$ are better

The approximation can be poor because

- Poor optimization
- The family $\Omega_{\mathcal{Q}}$ isn't expressive enough

When does variational inference work?

When, in general, is $\overset{*}{\mathcal{Q}}(\theta)$ a good approximation for a given family $\Omega_{\mathcal{Q}}$?

It is hard to say.

Black-box variational inference

To perform VI, we need to solve

$$\overset{*}{\mathcal{Q}}(\theta) = \underset{\mathcal{Q}\in\Omega_{\mathcal{Q}}}{\operatorname{argmin}} \underbrace{\left(\underbrace{\underset{\mathcal{Q}(\theta)}{\mathbb{E}} \left[\log \mathcal{Q}(\theta)\right]}_{\text{Entropy of } \mathcal{Q}} - \underbrace{\underset{\mathcal{Q}(\theta)}{\mathbb{E}} \left[\log \mathcal{P}(\theta, y)\right]}_{\text{Often intractable}} - \overbrace{\mathcal{P}(y)}^{\text{Constant}} \right)}_{\mathrm{KL}(\mathcal{Q}(\theta)||\mathcal{P}(\theta))}.$$

How can we optimize this objective?

Black-box variational inference

To perform VI, we need to solve

$$\overset{*}{\mathcal{Q}}(\theta) = \underset{\mathcal{Q}\in\Omega_{\mathcal{Q}}}{\mathrm{argmin}} \underbrace{\left(\underbrace{\underset{\mathcal{Q}(\theta)}{\mathbb{E}}\left[\log \mathcal{Q}(\theta)\right]}_{\text{Entropy of } \mathcal{Q}} - \underbrace{\underset{\mathcal{Q}(\theta)}{\mathbb{E}}\left[\log \mathcal{P}(\theta, y)\right]}_{\text{Often intractable}} - \overbrace{\mathcal{P}(y)}^{\text{Constant}} \right)}_{\mathrm{KL}(\mathcal{Q}(\theta)||\mathcal{P}(\theta))}.$$

How can we optimize this objective? **Black-box VI [Ranganath et al., 2014]:**

- Parameterize the family $\Omega_{\mathcal{Q}}$ using $\eta \in \mathbb{R}^{D_\eta}$ (so we have $\mathcal{Q}(\theta|\eta)$)
  - We will study **ADVI**, which takes $\mathcal{Q}(\theta|\eta)$ to be Gaussian [Kucukelbir et al., 2017].
  - The parameters η are the means and covariance ("mean-field" or "full-rank")
- Re-write the objective (using the reparameterization trick) as

$$\underset{\eta}{\mathrm{argmin}}\, F(\eta) \quad \text{where} \quad F(\eta) := \underset{\mathcal{N}_{\mathrm{std}}(z)}{\mathbb{E}}\left[f(\eta, z)\right].$$

- Use autodiff to differentiate $\eta \mapsto f(\eta, z)$
- Optimize with stochastic optimization using draws $z_n \sim \mathcal{N}_{\mathrm{std}}(z)$.

Black-box variational inference

To perform VI, we need to solve

$$\overset{*}{\mathcal{Q}}(\theta) = \underset{\mathcal{Q} \in \Omega_{\mathcal{Q}}}{\operatorname{argmin}} \underbrace{\left(\underbrace{\underset{\mathcal{Q}(\theta)}{\mathbb{E}} [\log \mathcal{Q}(\theta)]}_{\text{Entropy of } \mathcal{Q}} - \underbrace{\underset{\mathcal{Q}(\theta)}{\mathbb{E}} [\log \mathcal{P}(\theta, y)]}_{\text{Often intractable}} - \overbrace{\mathcal{P}(y)}^{\text{Constant}} \right)}_{\mathrm{KL}(\mathcal{Q}(\theta)||\mathcal{P}(\theta))}.$$

How can we optimize this objective? **Black-box VI [Ranganath et al., 2014]:**

- Parameterize the family $\Omega_{\mathcal{Q}}$ using $\eta \in \mathbb{R}^{D_\eta}$ (so we have $\mathcal{Q}(\theta|\eta)$)
  - We will study **ADVI**, which takes $\mathcal{Q}(\theta|\eta)$ to be Gaussian [Kucukelbir et al., 2017].
  - The parameters η are the means and covariance ("mean-field" or "full-rank")
- Re-write the objective (using the reparameterization trick) as

$$\underset{\eta}{\operatorname{argmin}} F(\eta) \quad \text{where} \quad F(\eta) := \underset{\mathcal{N}_{\mathrm{std}}(z)}{\mathbb{E}} [f(\eta, z)].$$

- Use autodiff to differentiate $\eta \mapsto f(\eta, z)$
- Optimize with stochastic optimization using draws $z_n \sim \mathcal{N}_{\mathrm{std}}(z)$.

We propose:

Instead of $\underset{\mathcal{N}_{\mathrm{std}}(z)}{\mathbb{E}} [f(\eta, z)]$, optimize $\frac{1}{N} \sum_{n=1}^{N} f(\eta, z_n)$ for fixed $z_n \overset{iid}{\sim} \mathcal{N}_{\mathrm{std}}(z)$.

Two approaches

$$\text{Consider} \quad \operatorname*{argmin}_{\eta} F(\eta) \quad \text{where} \quad F(\eta) := \mathop{\mathbb{E}}_{\mathcal{N}_{\mathrm{std}}(z)} \left[f(\eta, z) \right].$$

Let $\mathcal{Z}_N = \{z_1, \dots, z_N\} \overset{iid}{\sim} \mathcal{N}_{\mathrm{std}}(z)$, and let $\hat{F}(\eta|\mathcal{Z}_N) := \frac{1}{N}\sum_{n=1}^{N} f(\eta, z_n)$.

Two approaches

$$\text{Consider} \quad \underset{\eta}{\operatorname{argmin}} F(\eta) \quad \text{where} \quad F(\eta) := \underset{\mathcal{N}_{\mathrm{std}}(z)}{\mathbb{E}} \left[f(\eta, z) \right].$$

Let $\mathcal{Z}_N = \{z_1, \ldots, z_N\} \overset{iid}{\sim} \mathcal{N}_{\mathrm{std}}(z)$, and let $\hat{F}(\eta|\mathcal{Z}_N) := \frac{1}{N} \sum_{n=1}^{N} f(\eta, z_n)$.

Algorithm 1

Stochastic gradient (SG)

ADVI (and most BBVI)

Fix N (typically $N = 1$)
$t \leftarrow 0$
while Not converged **do**
  $t \leftarrow t + 1$
  Draw $\mathcal{Z}_N$
  $\Delta_S \leftarrow \nabla_\eta \hat{F}(\eta_{t-1}|\mathcal{Z}_N)$
  $\alpha_t \leftarrow \mathrm{SetStepSize}(\text{Past state})$
  $\eta_t \leftarrow \eta_{t-1} - \alpha_t \Delta_S$
  $\mathrm{AssessConvergence}(\text{Past state})$
end while
return η_t or $\frac{1}{M} \sum_{t'=t-M}^{t} \eta_{t'}$

Two approaches

$$\text{Consider} \quad \underset{\eta}{\operatorname{argmin}} F(\eta) \quad \text{where} \quad F(\eta) := \underset{\mathcal{N}_{\mathrm{std}}(z)}{\mathbb{E}} [f(\eta, z)].$$

Let $\mathcal{Z}_N = \{z_1, \ldots, z_N\} \overset{iid}{\sim} \mathcal{N}_{\mathrm{std}}(z)$, and let $\hat{F}(\eta|\mathcal{Z}_N) := \frac{1}{N}\sum_{n=1}^{N} f(\eta, z_n)$.

Algorithm 1
Stochastic gradient (SG)
ADVI (and most BBVI)

- Fix N (typically $N = 1$)
- $t \leftarrow 0$
- **while** Not converged **do**
  - $t \leftarrow t + 1$
  - Draw $\mathcal{Z}_N$
  - $\Delta_S \leftarrow \nabla_\eta \hat{F}(\eta_{t-1}|\mathcal{Z}_N)$
  - $\alpha_t \leftarrow \mathrm{SetStepSize}(\text{Past state})$
  - $\eta_t \leftarrow \eta_{t-1} - \alpha_t \Delta_S$
  - $\mathrm{AssessConvergence}(\text{Past state})$
- **end while**
- **return** η_t or $\frac{1}{M}\sum_{t'=t-M}^{t} \eta_{t'}$

Algorithm 2
Sample average approximation (SAA)
Deterministic ADVI (DADVI) (proposal)

- Fix N (our experiments use $N = 30$)
- Draw $\mathcal{Z}_N$
- $t \leftarrow 0$
- **while** Not converged **do**
  - $t \leftarrow t + 1$
  - $\Delta_D \leftarrow \mathrm{GetStep}(\hat{F}(\cdot|\mathcal{Z}_N), \eta_{t-1})$
  - $\eta_t \leftarrow \eta_{t-1} + \Delta_D$
  - $\mathrm{AssessConvergence}(\hat{F}(\cdot|\mathcal{Z}_N), \eta_t)$
- **end while**
- **return** η_t

Two approaches

$$\text{Consider} \quad \underset{\eta}{\operatorname{argmin}} F(\eta) \quad \text{where} \quad F(\eta) := \underset{\mathcal{N}_{\mathrm{std}}(z)}{\mathbb{E}} \left[f(\eta, z) \right].$$

Let $\mathcal{Z}_N = \{z_1, \dots, z_N\} \overset{iid}{\sim} \mathcal{N}_{\mathrm{std}}(z)$, and let $\hat{F}(\eta|\mathcal{Z}_N) := \frac{1}{N} \sum_{n=1}^{N} f(\eta, z_n)$.

Algorithm 1
Stochastic gradient (SG)
ADVI (and most BBVI)

- Fix N (typically $N = 1$)
- $t \leftarrow 0$
- **while** Not converged **do**
  - $t \leftarrow t + 1$
  - Draw $\mathcal{Z}_N$
  - $\Delta_S \leftarrow \nabla_\eta \hat{F}(\eta_{t-1}|\mathcal{Z}_N)$
  - $\alpha_t \leftarrow \mathrm{SetStepSize}(\mathrm{Past\ state})$
  - $\eta_t \leftarrow \eta_{t-1} - \alpha_t \Delta_S$
  - $\mathrm{AssessConvergence}(\mathrm{Past\ state})$
- **end while**
- **return** η_t or $\frac{1}{M} \sum_{t'=t-M}^{t} \eta_{t'}$

Algorithm 2
Sample average approximation (SAA)
Deterministic ADVI (DADVI) (proposal)

- Fix N (our experiments use $N = 30$)
- Draw $\mathcal{Z}_N$
- $t \leftarrow 0$
- **while** Not converged **do**
  - $t \leftarrow t + 1$
  - $\Delta_D \leftarrow \mathrm{GetStep}(\hat{F}(\cdot|\mathcal{Z}_N), \eta_{t-1})$
  - $\eta_t \leftarrow \eta_{t-1} + \Delta_D$
  - $\mathrm{AssessConvergence}(\hat{F}(\cdot|\mathcal{Z}_N), \eta_t)$
- **end while**
- **return** η_t

Our proposal: Apply Algorithm 2 with the ADVI objective.
Take better steps, easily assess convergence, with less tuning.

Experiments

For each of a range of models (next slide), we compared:

- **NUTS:** The "no-U-turn" MCMC sampler as implemented by PyMC [Salvatier et al., 2016]. We used this as the "ground truth" posterior.

Experiments

For each of a range of models (next slide), we compared:

- **NUTS:** The "no-U-turn" MCMC sampler as implemented by PyMC [Salvatier et al., 2016]. We used this as the "ground truth" posterior.

- **DADVI:** We used $N = 30$ draws for DADVI for each model. We optimized using an off-the-shelf second-order Newton trust region method (`trust-ncg` in `scipy.optimize.minimize`) with no tuning or preconditioning.

Experiments

For each of a range of models (next slide), we compared:

- **NUTS:** The "no-U-turn" MCMC sampler as implemented by PyMC [Salvatier et al., 2016]. We used this as the "ground truth" posterior.

- **DADVI:** We used $N = 30$ draws for DADVI for each model. We optimized using an off-the-shelf second-order Newton trust region method (`trust-ncg` in `scipy.optimize.minimize`) with no tuning or preconditioning.

Stochastic ADVI methods:

- Mean field ADVI: We used the PyMC implementation of ADVI, together with its default termination criterion (based on parameter differences).
- Full-rank ADVI: We used the PyMC implementation of full-rank ADVI, together with the default termination criterion for ADVI described above.
- RAABBVI: To run RAABBVI, we used the public package `viabel`, provided by Welandawe et al. [2022].

We terminated unconverged stochastic ADVI after 100,000 iterations.

We evaluated each method on a range of models.

Model Name	Dim D_θ	NUTS runtime	Description
ARM (53 models)	Median 5 (max 176)	median 39 seconds (max 16 minutes)	A range of linear models, GLMs, and GLMMs
Microcredit	124	597 minutes	Hierarchical model with heavy tails and zero inflation
Occupancy	$1,884$	251 minutes	Binary regression with highly crossed random effects
Tennis	$5,014$	57 minutes	Binary regression with highly crossed random effects
POTUS	$15,098$	643 minutes	Autoregressive time series with random effects

Table 1: Model summaries.

Comparisons

To form a common scale for the accuracy of the posteriors, we report:

$$\varepsilon^{\mu}_{\text{METHOD}} := \frac{\mu_{\text{METHOD}} - \mu_{\text{NUTS}}}{\sigma_{\text{NUTS}}} \qquad \varepsilon^{\sigma}_{\text{METHOD}} := \frac{\sigma_{\text{METHOD}} - \sigma_{\text{NUTS}}}{\sigma_{\text{NUTS}}}.$$

where

$$\mu_{\text{METHOD}} := \textsc{method} \text{ posterior mean} \qquad \sigma_{\text{METHOD}} := \textsc{method} \text{ posterior SD}.$$

Comparisons

To form a common scale for the accuracy of the posteriors, we report:

$$\varepsilon^{\mu}_{\textsc{method}} := \frac{\mu_{\textsc{method}} - \mu_{\textsc{nuts}}}{\sigma_{\textsc{nuts}}} \qquad \varepsilon^{\sigma}_{\textsc{method}} := \frac{\sigma_{\textsc{method}} - \sigma_{\textsc{nuts}}}{\sigma_{\textsc{nuts}}}.$$

where

$$\mu_{\textsc{method}} := \textsc{method} \text{ posterior mean} \qquad \sigma_{\textsc{method}} := \textsc{method} \text{ posterior SD.}$$

We measure computational cost using both

- **Wall time** and
- **Number of model evaluations** (gradients, Hessian-vector products).

We compare achieved objective values using a large number of independent samples.

We report objective values and computation cost relative to DADVI.

Posterior mean accuracy

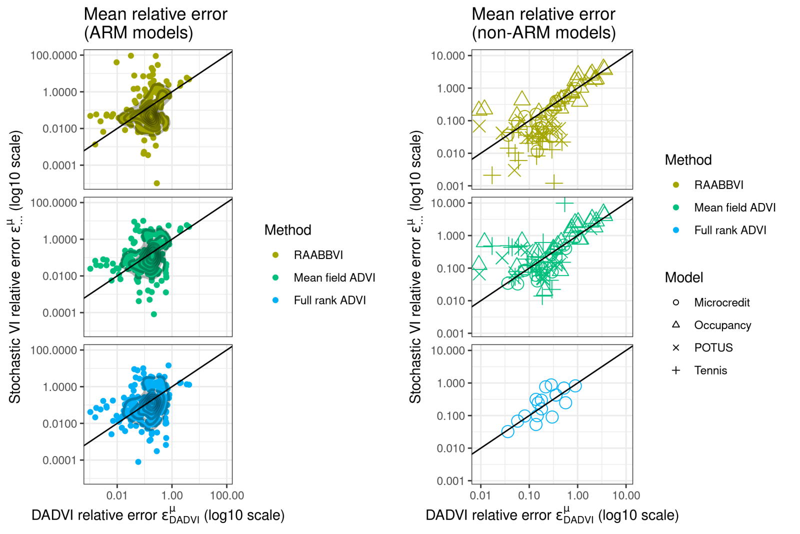


Figure 1: Posterior mean accuracy (relative to MCMC posterior standard deviation). Each point is a single named parameter in a single model. Points above the diagonal line indicate better DADVI or LRVB performance.

Computational cost for ARM models

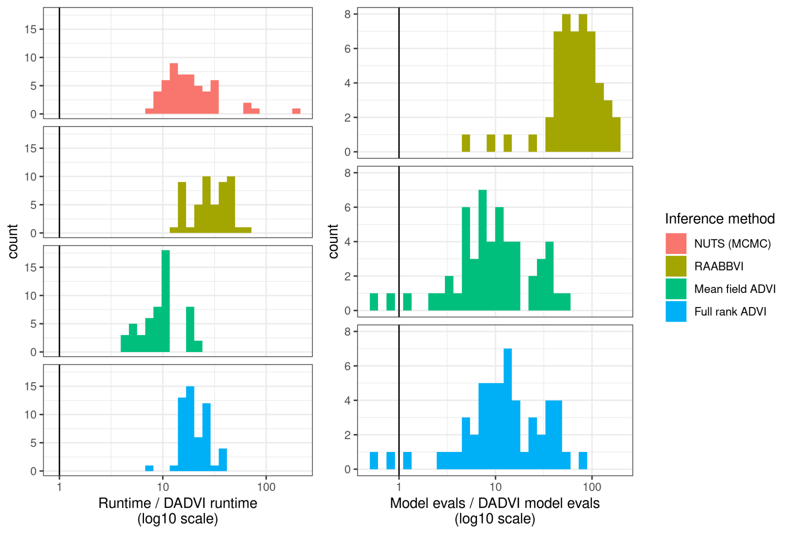


Figure 2: Runtimes and model evaluation counts for the ARM models. Results are reported divided by the corresponding value for DADVI.

Computational cost for non-ARM models

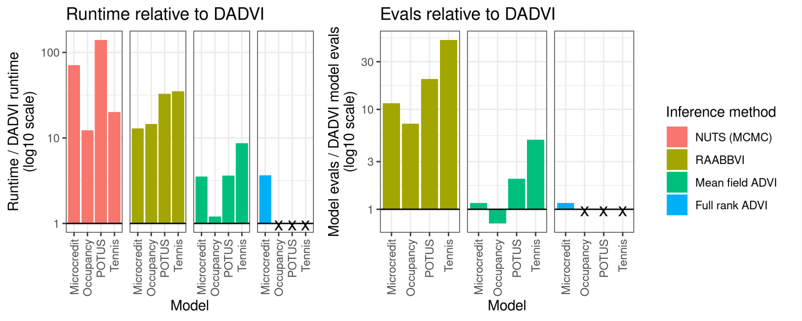


Figure 3: Runtimes and model evaluation counts for the non-ARM models. Results are reported divided by the corresponding value for DADVI. Missing model / method combinations are marked with an X.

Optimization traces for ARM models

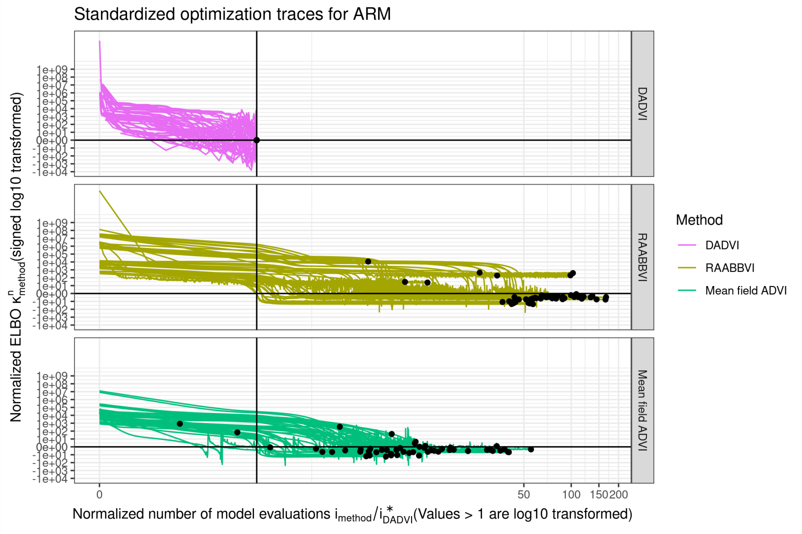


Figure 4: Optimization traces for the ARM models. Black dots show the termination point of each method. Dots above the horizontal black line mean that DADVI found a better ELBO. Dots to the right of the black line mean that DADVI terminated sooner in terms of model evaluations.

Optimization traces for non-ARM models

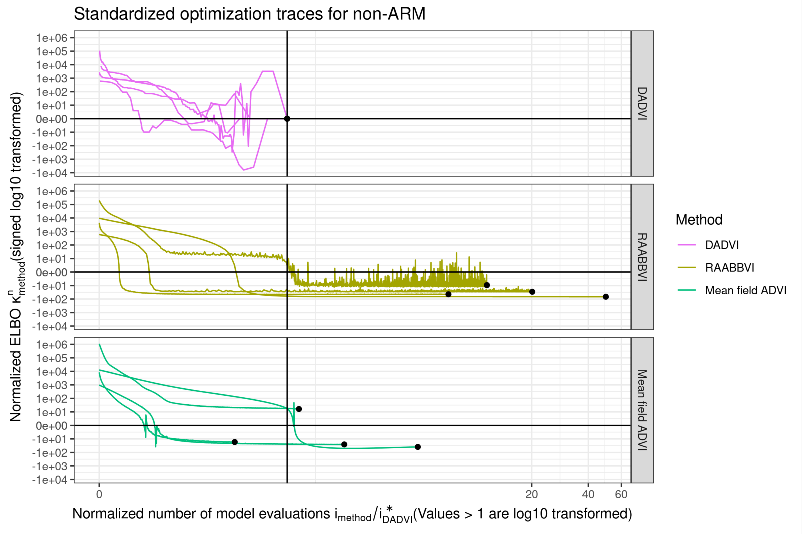


Figure 5: Traces for non-ARM models. Black dots show the termination point of each method. Dots above the horizontal black line mean that DADVI found a better ELBO. Dots to the right of the black line mean that DADVI terminated sooner in terms of model evaluations.

Experiment summary

$\Rightarrow$ DADVI is faster, simpler, and the posterior means are not worse.

But DADVI can additionally provide:

- Simple estimates of approximation error
- Improved (LR) posterior covariance esimates

Linear response covariances and sampling uncertainty

Intractable objective:

$$\overset{*}{\eta} = \underset{\eta \in \mathbb{R}^{D_\eta}}{\mathrm{argmin}} \underset{\mathcal{N}_{\mathrm{std}}(z)}{\mathbb{E}} \left[f(\eta, z) \right]$$

DADVI approximation:

$$\hat{\eta}(\mathcal{Z}_N) = \underset{\eta \in \mathbb{R}^{D_\eta}}{\mathrm{argmin}} \frac{1}{N} \sum_{n=1}^{N} f(\eta, z_n).$$

What is the error of the DADVI approximation $\hat{\eta} - \overset{*}{\eta}$?

Linear response covariances and sampling uncertainty

Intractable objective:

$$\overset{*}{\eta} = \underset{\eta \in \mathbb{R}^{D_\eta}}{\operatorname{argmin}} \underset{\mathcal{N}_{\mathrm{std}}(z)}{\mathbb{E}} \left[f(\eta, z) \right]$$

DADVI approximation:

$$\hat{\eta}(\mathcal{Z}_N) = \underset{\eta \in \mathbb{R}^{D_\eta}}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^{N} f(\eta, z_n).$$

What is the error of the DADVI approximation $\hat{\eta} - \overset{*}{\eta}$?

$\Leftrightarrow$ What is the distribution of the DADVI error $\hat{\eta} - \overset{*}{\eta}$ under sampling of $\mathcal{Z}_N$?

Answer: The same as a that of any M-estimator: asymptotically normal (as N grows)

Linear response covariances and sampling uncertainty

Intractable objective:

$$\overset{*}{\eta} = \underset{\eta \in \mathbb{R}^{D_\eta}}{\mathrm{argmin}} \underset{\mathcal{N}_{\mathrm{std}}(z)}{\mathbb{E}} [f(\eta, z)]$$

DADVI approximation:

$$\hat{\eta}(\mathcal{Z}_N) = \underset{\eta \in \mathbb{R}^{D_\eta}}{\mathrm{argmin}} \frac{1}{N} \sum_{n=1}^{N} f(\eta, z_n).$$

What is the error of the DADVI approximation $\hat{\eta} - \overset{*}{\eta}$?

$\Leftrightarrow$ What is the distribution of the DADVI error $\hat{\eta} - \overset{*}{\eta}$ under sampling of $\mathcal{Z}_N$?

Answer: The same as a that of any M-estimator: asymptotically normal (as N grows)

Posterior variances are often badly estimated by mean-field (MF) approximations.

Linear response (LR) covariances improve covariance estimates by computing *sensitivity* of the variational means to particular perturbations. [Giordano et al., 2018]

Example: With a correlated Gaussian $\mathcal{P}(\theta|y)$, the ADVI means are exactly correct, the ADVI variances are underestimated, and LR covariances are exactly correct.

Linear response covariances and sampling uncertainty

Intractable objective:

$$\overset{*}{\eta} = \underset{\eta \in \mathbb{R}^{D_\eta}}{\mathrm{argmin}} \underset{\mathcal{N}_{\mathrm{std}}(z)}{\mathbb{E}} \left[f(\eta, z) \right]$$

DADVI approximation:

$$\hat{\eta}(\mathcal{Z}_N) = \underset{\eta \in \mathbb{R}^{D_\eta}}{\mathrm{argmin}} \frac{1}{N} \sum_{n=1}^{N} f(\eta, z_n).$$

What is the error of the DADVI approximation $\hat{\eta} - \overset{*}{\eta}$?

$\Leftrightarrow$ What is the distribution of the DADVI error $\hat{\eta} - \overset{*}{\eta}$ under sampling of $\mathcal{Z}_N$?

Answer: The same as a that of any M-estimator: asymptotically normal (as N grows)

Posterior variances are often badly estimated by mean-field (MF) approximations.

Linear response (LR) covariances improve covariance estimates by computing *sensitivity* of the variational means to particular perturbations. [Giordano et al., 2018]

Example: With a correlated Gaussian $\mathcal{P}(\theta|y)$, the ADVI means are exactly correct, the ADVI variances are underestimated, and LR covariances are exactly correct.

Both DADVI error and LR covariances can be computed from the DADVI objective.

Stochastic ADVI does not produce an actual optimum of any tractable objective, so LR and M-estimator computations are unavailable.

Posterior standard deviation accuracy

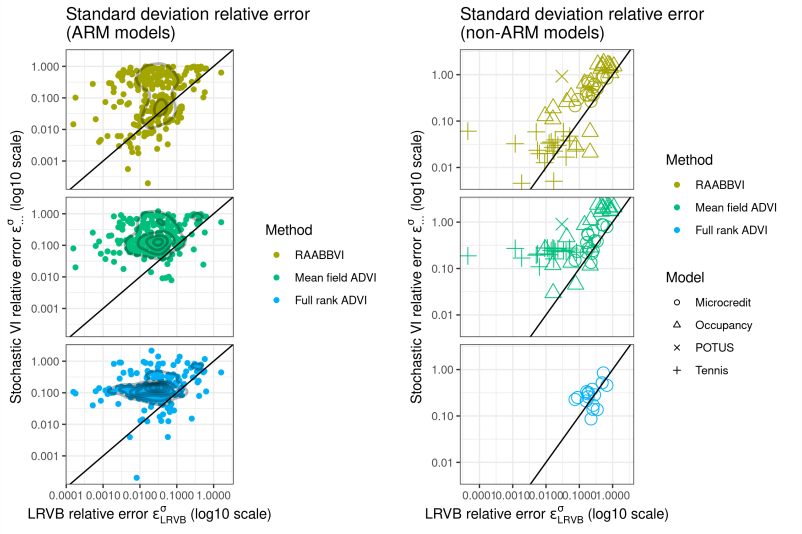


Figure 6: Posterior sd relative accuracy. Each point is a single named parameter in a single model. Points above the diagonal line indicate better DADVI or LRVB performance.

DADVI approximation error accuracy

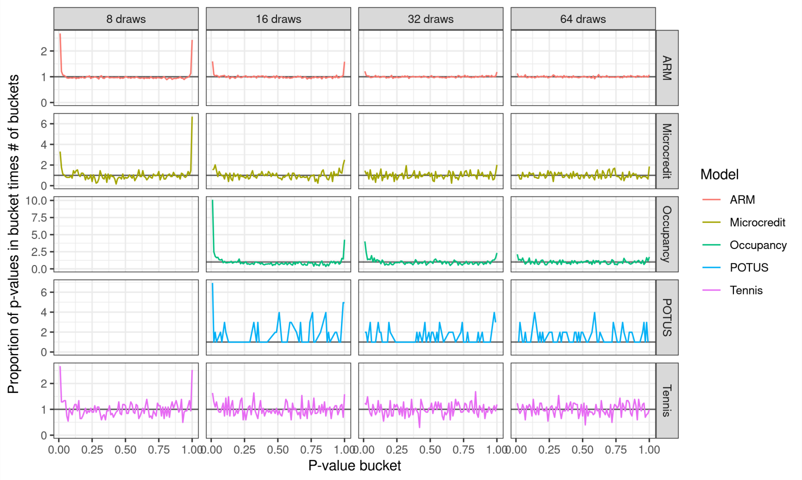


Figure 7: Density estimates of $\Phi(\varepsilon^{\xi})$ for difference models. All the ARM models are grouped together for ease of visualization. Each panel shows a binned estimate of the density of $\Phi(\varepsilon^{\xi})$ for a particular model and number of draws N. Values close to one (a uniform density) indicate good frequentist performance. CG failed for the Occupancy and POTUS models with only 8 draws, possibly indicating poor optimization performance with so few samples.

Previous theoretical results

Intractable objective:

$$\overset{*}{\eta} = \underset{\eta \in \mathbb{R}^{D_\eta}}{\mathrm{argmin}} \underset{\mathcal{N}_{\mathrm{std}}(z)}{\mathbb{E}} [f(\eta, z)]$$

SAA approximation (DADVI):

$$\hat{\eta}(\mathcal{Z}_N) = \underset{\eta \in \mathbb{R}^{D_\eta}}{\mathrm{argmin}} \frac{1}{N} \sum_{n=1}^{N} f(\eta, z_n).$$

The idea of optimizing $\hat{F}$ instead of SG on F is old and well-studied in the optimization literature, where $\hat{F}$ is known as the **Sample average approximation (SAA)**.

Yet SAA is rarely used for BBVI.[1] One possible reason is the following:

Theorem [Nemirovski et al., 2009]: In general, the error of both SG and SAA scale as $\sqrt{D_\theta / N}$, where, for SG, N is the *total number of samples used*.

[1] Some exceptions I'm aware of: Giordano et al. [2018, 2022], Wycoff et al. [2022], Burroni et al. [2023].

Previous theoretical results

Intractable objective:

$$\overset{*}{\eta} = \underset{\eta \in \mathbb{R}^{D_\eta}}{\operatorname{argmin}} \underset{\mathcal{N}_{\mathrm{std}}(z)}{\mathbb{E}} [f(\eta, z)]$$

SAA approximation (DADVI):

$$\hat{\eta}(\mathcal{Z}_N) = \underset{\eta \in \mathbb{R}^{D_\eta}}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^{N} f(\eta, z_n).$$

The idea of optimizing $\hat{F}$ instead of SG on F is old and well-studied in the optimization literature, where $\hat{F}$ is known as the **Sample average approximation (SAA)**.

Yet SAA is rarely used for BBVI.[1] One possible reason is the following:

Theorem [Nemirovski et al., 2009]: In general, the error of both SG and SAA scale as $\sqrt{D_\theta / N}$, where, for SG, N is the *total number of samples used*.

- For SG, each z_n gets used once (for a single gradient step)
- For SAA, each z_n gets used once per optimization step (of which the are many).
- Often, in higher dimensions, SAA requires more optimization steps.

Corollary: [Kim et al., 2015] In general, for a given accuracy, the computation required for SAA scales worse than SG as the dimension D_θ grows.

[1] Some exceptions I'm aware of: Giordano et al. [2018, 2022], Wycoff et al. [2022], Burroni et al. [2023].

Previous theoretical results

Intractable objective:

$$\overset{*}{\eta} = \underset{\eta \in \mathbb{R}^{D_\eta}}{\mathrm{argmin}} \underset{\mathcal{N}_{\mathrm{std}}(z)}{\mathbb{E}} [f(\eta, z)]$$

SAA approximation (DADVI):

$$\hat{\eta}(\mathcal{Z}_N) = \underset{\eta \in \mathbb{R}^{D_\eta}}{\mathrm{argmin}} \frac{1}{N} \sum_{n=1}^{N} f(\eta, z_n).$$

The idea of optimizing $\hat{F}$ instead of SG on F is old and well-studied in the optimization literature, where $\hat{F}$ is known as the **Sample average approximation (SAA)**.

Yet SAA is rarely used for BBVI.[1] One possible reason is the following:

Theorem [Nemirovski et al., 2009]: In general, the error of both SG and SAA scale as $\sqrt{D_\theta / N}$, where, for SG, N is the *total number of samples used*.

- For SG, each z_n gets used once (for a single gradient step)
- For SAA, each z_n gets used once per optimization step (of which the are many).
- Often, in higher dimensions, SAA requires more optimization steps.

Corollary: [Kim et al., 2015] In general, for a given accuracy, the computation required for SAA scales worse than SG as the dimension D_θ grows.

But we got good results with D_θ as high as $15,098$ using only only $N = 30$. Why?

[1] Some exceptions I'm aware of: Giordano et al. [2018, 2022], Wycoff et al. [2022], Burroni et al. [2023].

Some first steps

Theorem [Giordano et al., 2023]: When $\mathcal{P}(\theta|y)$ is multivariate normal, and we use the mean-field Gaussian approximation, then, for any particular entry η_d of η, then $\left|\hat{\eta}_d - \overset{*}{\eta}_d\right| = O_p(N^{-1/2})$ irrespective of D_θ.

Some first steps

Theorem [Giordano et al., 2023]: When $\mathcal{P}(\theta|y)$ is multivariate normal, and we use the mean-field Gaussian approximation, then, for any particular entry η_d of η, then $\left|\hat{\eta}_d - \overset{*}{\eta}_d\right| = O_p(N^{-1/2})$ irrespective of D_θ.

Theorem [Giordano et al., 2023]: Assume $\mathcal{P}(\theta|y)$ has a "global-local" structure:

$$\theta = (\gamma, \lambda_1, \ldots, \lambda_{D_\lambda}) \qquad \mathcal{P}(\gamma, \lambda_1, \ldots, \lambda_{D_\lambda}|y) = \prod_{d=1}^{D_\lambda} \mathcal{P}(\gamma, \lambda_d|y).$$

Assume that the dimension of γ and each λ_d stays fixed as D_λ grows.

Under regularity conditions, the DADVI error scales as $\sqrt{\log D_\lambda / N}$, not $\sqrt{D_\lambda / N}$.

Some first steps

Theorem [Giordano et al., 2023]: When $\mathcal{P}(\theta|y)$ is multivariate normal, and we use the mean-field Gaussian approximation, then, for any particular entry η_d of η, then $\left|\hat{\eta}_d - \overset{*}{\eta}_d\right| = O_p(N^{-1/2})$ irrespective of D_θ.

Theorem [Giordano et al., 2023]: Assume $\mathcal{P}(\theta|y)$ has a "global-local" structure:

$$\theta = (\gamma, \lambda_1, \ldots, \lambda_{D_\lambda}) \qquad \mathcal{P}(\gamma, \lambda_1, \ldots, \lambda_{D_\lambda}|y) = \prod_{d=1}^{D_\lambda} \mathcal{P}(\gamma, \lambda_d|y).$$

Assume that the dimension of γ and each λ_d stays fixed as D_λ grows.

Under regularity conditions, the DADVI error scales as $\sqrt{\log D_\lambda / N}$, not $\sqrt{D_\lambda / N}$.

Proposal: The "in general" analysis of [Nemirovski et al., 2009] is too general for many practically interesting BBVI problems.

A negative result for expressive approximations

Theorem [Giordano et al., 2023]: Assume that $N < D_\theta$, and that we use a full-rank Gaussian approximation. Then the DADVI objective is unbounded below, and optimization of the DADVI objective will approach a degenerate point mass at $\mathrm{argmax}_\theta \log \mathcal{P}(\theta|y)$.

A negative result for expressive approximations

Theorem [Giordano et al., 2023]: Assume that $N < D_\theta$, and that we use a full-rank Gaussian approximation. Then the DADVI objective is unbounded below, and optimization of the DADVI objective will approach a degenerate point mass at $\mathrm{argmax}_\theta \log \mathcal{P}(\theta|y)$.

Proof sketch: For any value of the variational mean, the DADVI objective only depends on $\mathcal{P}(\theta|y)$ evaluated in a subspace spanned by $\mathcal{Z}_N$. The variational objective can be driven to $-\infty$ by driving the variance to zero in the subspace orthogonal to $\mathcal{Z}_N$.

A negative result for expressive approximations

Theorem [Giordano et al., 2023]: Assume that $N < D_\theta$, and that we use a full-rank Gaussian approximation. Then the DADVI objective is unbounded below, and optimization of the DADVI objective will approach a degenerate point mass at $\mathrm{argmax}_\theta \log \mathcal{P}(\theta|y)$.

Proof sketch: For any value of the variational mean, the DADVI objective only depends on $\mathcal{P}(\theta|y)$ evaluated in a subspace spanned by $\mathcal{Z}_N$. The variational objective can be driven to $-\infty$ by driving the variance to zero in the subspace orthogonal to $\mathcal{Z}_N$.

Proposal: All sufficiently expressive variational approximations (e.g. normalizing flows) will fail in the same way in high dimensions. However, this pathology can be obscured and overlooked in practice by low-quality optimization.

Conclusion

Black Box Variational Inference with a Deterministic Objective: Faster, More Accurate, and Even More Black Box.

Giordano, R.*, Ingram, M.*, Broderick, T. (* joint first authors), 2023.

(Arxiv preprint here.)

- By fixing the randomness in the ADVI objective, DADVI provides BBVI that is easier to use, faster, and more accurate than stochastic gradient.
- The approximation used by DADVI will not work in high dimensions for sufficiently expressive approximating distributions (e.g., full-rank ADVI).
- There appears to be a gap between the optimization literature and BBVI practice in high dimensions for a class of practically interesting problems.

References

C. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006. Chapter 10.

D. Blei, A. Kucukelbir, and J. McAuliffe. Variational inference: A review for statisticians. *arXiv preprint arXiv:1601.00670*, 2016.

J. Burroni, J. Domke, and D. Sheldon. Sample average approximation for Black-Box VI. *arXiv preprint arXiv:2304.06803*, 2023.

R. Giordano, R. Liu, M. I. Jordan, and T. Broderick. Evaluating sensitivity to the stick-breaking prior in Bayesian nonparametrics. *Bayesian Analysis*, 1(1):1–34, 2022.

R. Giordano, M. Ingram, and T. Broderick. Black box variational inference with a deterministic objective: Faster, more accurate, and even more black box. *arXiv preprint arXiv:2304.05527*, 2023.

T. Giordano, T. Broderick, and M. I. Jordan. Covariances, Robustness, and Variational Bayes. *Journal of Machine Learning Research*, 19(51):1–49, 2018. URL `http://jmlr.org/papers/v19/17-670.html`.

S. Kim, R. Pasupathy, and S. Henderson. A guide to sample average approximation. *Handbook of simulation optimization*, pages 207–243, 2015.

A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. Blei. Automatic Differentiation Variational Inference. *Journal of Machine Learning Research*, 18(14):1–45, 2017. URL `http://jmlr.org/papers/v18/16-107.html`.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black Box Variational Inference. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 814–822, Reykjavik, Iceland, 22–25 Apr 2014. PMLR. URL `http://proceedings.mlr.press/v33/ranganath14.html`.

J. Salvatier, T. Wiecki, and C. Fonnesbeck. Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, 2:e55, apr 2016. doi: 10.7717/peerj-cs.55. URL `https://doi.org/10.7717/peerj-cs.55`.

M. Welandawe, M. Andersen, A. Vehtari, and J. Huggins. Robust, automated, and accurate black-box variational inference. *arXiv preprint arXiv:2203.15945*, 2022.

N. Wycoff, A. Arab, K. Donato, and L. Singh. Sparse bayesian lasso via a variable-coefficient ℓ_1 penalty. *arXiv preprint arXiv:2211.05089*, 2022.

Supplemental material

Linear response covariances

Posterior variances are often badly estimated by mean-field (MF) approximations.

Example: With a correlated Gaussian $\mathcal{P}(\theta|y)$ with ADVI, the ADVI means are correct, but the ADVI variances are underestimated.

Take a variational approximation $\overset{*}{\eta} := \mathrm{argmin}_{\eta \in \mathbb{R}^{D_\eta}} \mathrm{KL}_{\mathrm{VI}}(\eta)$. Often,

$$\underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathbb{E}}[\theta] \approx \underset{\mathcal{P}(\theta|y)}{\mathbb{E}}[\theta] \quad \text{but} \quad \underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathrm{Var}}(\theta) \neq \underset{\mathcal{P}(\theta|y)}{\mathrm{Var}}(\theta). \tag{1}$$

Example: Correlated Gaussian $\mathcal{P}(\theta|y)$ with ADVI.

Linear response covariances use the fact that, if $\mathcal{P}(\theta|y,t) \propto \mathcal{P}(\theta|y)\exp(t\theta)$, then

$$\left.\frac{d \underset{\mathcal{P}(\theta|y,t)}{\mathbb{E}}[\theta]}{dt}\right|_{t=0} = \underset{\mathcal{P}(\theta|y)}{\mathrm{Cov}}(\theta). \tag{2}$$

Let $\overset{*}{\eta}(t)$ be the variational approximation to $\mathcal{P}(\theta|y,t)$, and take

$$\underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathrm{LRCov}}(\theta) = \left.\frac{d \underset{\mathcal{Q}(\theta|\overset{*}{\eta}(t))}{\mathbb{E}}[\theta]}{dt}\right|_{t=0} = \left(\nabla_\eta \underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathbb{E}}[\theta]\right) \left(\nabla^2_\eta \mathrm{KL}_{\mathrm{VI}}\left(\overset{*}{\eta}\right)\right)^{-1} \left(\nabla_\eta \underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathbb{E}}[\theta]\right)$$

Example: For ADVI with a correlated Gaussian $\mathcal{P}(\theta|y)$, $\underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathrm{LRCov}}(\theta) = \underset{\mathcal{Q}(\theta|\overset{*}{\eta})}{\mathrm{Cov}}(\theta)$.